DSTC 11 Track Proposal
Ambiguous Candidate Identification and Coreference Resolution for Immersive Multimodal Conversations

Meta AI & Meta Reality Labs

1 Motivation

The Situated and Interactive Multimodal Conversational AI (SIMMC) challenges [13, 11], held as part of DSTC9 [8] and DSTC10, pioneered the work for building the real-world assistant agents that can understand multimodal inputs (vision & conversations) and handle user requests. Throughout the two editions of the challenge, we provided two new benchmark datasets (SIMMC 1.0 and 2.0) for studying multimodal conversations with situated user context in the form of a co-observed image or virtual reality (VR) environment. Specifically, the SIMMC 2.0 dataset provided the assistant↔user task-oriented dialogs grounded on diverse photo-realistic VR renders of (synthetic) commercial stores with various referent objects, serving as a proxy for complex real-world scenarios. The earlier SIMMC challenge at DSTC10 saw a total of 16 model entries from participants around the world, establishing a new set of state-of-the-art baselines for the multimodal task-oriented dialog systems.

While the new SOTA models have drastically improved the performance on the previous benchmark tasks, several challenges remain in building the production-ready agent. One such challenge is the visual disambiguation which is often encountered in real-world multimodal conversations. For instance, the user ambiguously uses *'these two trousers'* in Fig. 1, which cannot be deterministically resolved. Thus, an assistant system needs to reason about this, identify the possible ambiguous candidates, and then ask a follow-up question to disambiguate. This is important to avoid a wrong resolution of such references leading to subsequent turns with false premises. Another advantage is that the assistant could insert implicit confirmation – "*Which of the two red jackets are you referring to - the one on the left or the one closer to you?*", as opposed to a generic response like: "*Which one are you referring to?*".

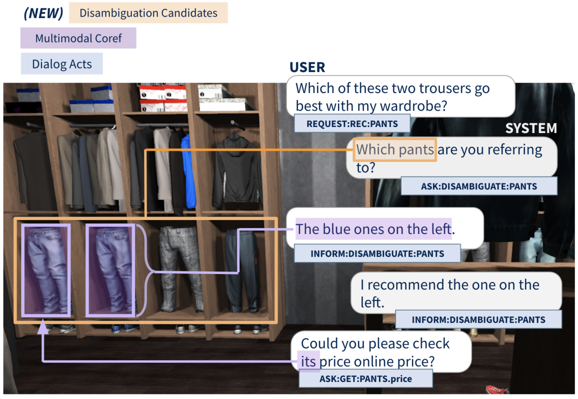
(NEW)
Disambiguation Candidates
Multimodal Coref
Dialog Acts
USER
Which of these two trousers go best with my wardrobe?
REQUEST:REC:PANTS
SYSTEM
Which pants are you referring to?
ASK:DISAMBIGUATE:PANTS
The blue ones on the left.
INFORM:DISAMBIGUATE:PANTS
I recommend the one on the left.
INFORM:DISAMBIGUATE:PANTS
Could you please check its price online price?
ASK:GET:PANTS.price

Figure 1: Illustration of a Situated Interactive Multimodal Conversation (SIMMC), which presents a task-oriented user↔assistant dialog grounded in a co-observed photorealistic multimodal context. The new version of the SIMMC dataset includes more fine-grained and precise annotations for referent disambiguation candidates, which poses new challenges for the Multimodal Coreference Resolution task (MM-Coref) and the Ambiguous Candidate Identification task.

To this end, we propose a third edition of the SIMMC challenge for the community to tackle and continue the effort towards building a successful multimodal assistant agent. In this edition of the challenge, we specifically focus on the key challenge of fine-grained visual disambiguation, which adds an important skill to assistant agents studied in the previous SIMMC challenge. To accommodate for this challenge, we provide the improved version of the of the dataset, SIMMC 2.1, where we augment the SIMMC 2.0 dataset with additional annotations (*i.e.* identification of all possible referent candidates given ambiguous mentions) and corresponding re-paraphrases to support the study and modeling of visual disambiguation (SIMMC 2.1). This new version of the dataset poses several interesting challenges such multimodal dialog state tracking given ambiguity, coreference resolutions ("*directly behind it*", "*grey jacket to the right of the one I mentioned*"), and disambiguation strategies ("*How much is that shirt*" → "*Which shirt are you referring to?*").

2 The Third SIMMC Challenge

We now detail the new version of the multimodal conversational dataset (SIMMC 2.1) (Sec. 2.2) and propose four main sub-tasks for the challenge (Tab. 1, Sec. 2.3).

2.1 Problem Setup

The SIMMC challenge studies the conversational scenarios where the virtual assistant shares a co-observed scene with a user. Specifically, the dataset targets the shopping experience as the domain of study, which often induces rich multimodal interactions around browsing visually grounded items in a physical store (fashion or furniture). The assistant agent is assumed to have access to the ground-truth meta information of every object in the scene, while users observe those objects only through the visual modality to describe and compose a request, as in the real-world applications. Each dialog in the dataset includes multiple viewpoints at different time steps throughout the session, corresponding to the scenarios where users are physically navigating the scene during the conversation. Therefore, the conversational models for the SIMMC challenge need to understand both user requests using both the dialog history and the state of the environment as multimodal context.

Note that the SIMMC problem setup where user and assistant co-observe the same scene allows for more natural multimodal coreferences to be used as part of user-assistant conversations. The previous literature in multimodal dialogs [1, 10, 4, 12, 6, 5] often assumes that dialog participants take the roles of a primary and secondary observer respectively, *i.e.* *questioner* and *answerer* similar to the Visual Question Answering [2] tasks, which does not address the real-world consumer scenario we are targetting. The SIMMC challenge also extends many of the key dialog tasks studied in the previous literature on conventional task-oriented dialog systems [9, 14, 3, 7] (*e.g.* DST, slot carryovers) to the unique multimodal settings.

2.2 The New SIMMC 2.1 Dataset

The SIMMC 2.1 dataset extends the original SIMMC 2.0 dataset [11] with additional annotations for fine-grained referent candidates and new utterances (details below).

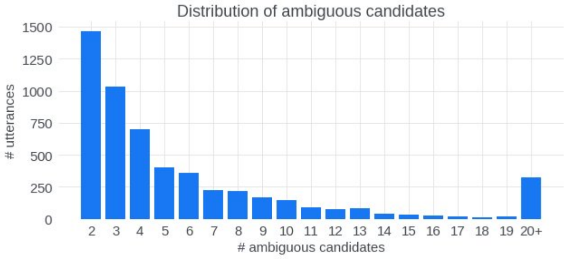

Figure 2: Distribution of ambiguous items.

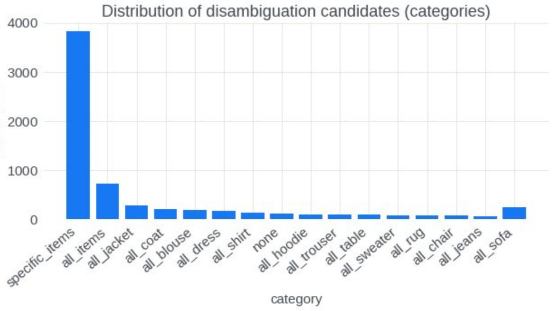

Figure 3: Distribution of ambiguous items per category.

Collection Process: The original dataset used the two-phase pipeline to collect dialogs (multimodal dialog simulation & manual human paraphrase), which can effectively collect natrual dialogs with the minimum the annotation overheads. Note that this approach extends the popular machine↔human collaborative dialog collection approaches [14, 15] to the unique multimodal settings. More details on the data collection approach can be found in [11].

Data Statistics: The dataset includes 11,244 dialogs (117,236 utterances), with the fine-grained ground-truth dialog labels (NLU/NLG/Coref) already in place. Table 2 shows the statistics of the dataset.

Additional Annotations in SIMMC 2.1. The key differences are (illustrated in Fig. 1). We collect candidates in the scene for turns with ambiguous references using human annotators. Further, we also ask the annotators to paraphrase the turn in case there is not enough ambiguity. This data collection process thus allows for richer coreferences, referential expressions, and disambiguation scenarios.

Annotation Analysis: We have $6.5k$ turns with ambiguous candidates, where the average number of candidates is 5.6. Fig. 2 and Fig. 3 represent the distribution of the candidates in these utterances per the number of items or per category.

Task Name	Goal	Evaluation
1. Ambiguous Candidate Identification	Given user utterances with ambiguous object mentions, resolve all referent candidate objects to their canonical ID(s) as defined by the catalog.	Object identification Precision / Recall / F1
2. Multimodal Coreference Resolution	Given user utterances with object mentions, resolve referent objects to their canonical ID(s) as defined by the catalog.	Coref Precision / Recall / F1
3. Multimodal Dialog State Tracking	Given user utterances, track user belief states across multiple turns.	Intent Accuracy, Slot Precision / Recall / F1
4. Response Generation	Given user utterances, ground-truth APIs and ground-truth object IDs, generate Assistant responses or retrieve from a candidate pool.	Generation: BLEU-4 score

Table 1: **Proposed tracks and descriptions.**

Total # dialogs	11,244
Total # utterances	117,236
Total # scenes	3,133
Avg # words per user turns	12
Avg # words per assistant turns	13.7
Avg # utterances per dialog	10.4
Avg # objects mentioned per dialog	4.7
Avg # objects in scene per dialog	19.7
Avg # candidates per ambiguous turn	5.6

Table 2: **SIMMC 2.1 Dataset Statistics**

Data format: The SIMMC 2.1 data will be provided in the same format as the earlier version of the datasets, making it easier for participants to use the various benchmark models publicly available [13, 11], or augment it with the previous version of the dataset for pretraining, etc. The raw pixel images of each scene as well as the pre-computed visual embeddings will be provided, allowing for easier adaptation for the NLP audience.

2.3 Challenge Tasks

We invite the DSTC community to build multimodal conversational agents for the following four benchmark tasks, addressing the key challenges in the multimodal conversational reasoning (summarized in Tab. 1). All the benchmark tasks require a strong computer vision capability as well as a multimodal conversational reasoning capability, to jointly process both the dialog and the visual contexts.

2.3.1 Ambiguous Candidates Identification

As a main focus of this edition of the challenge, this sub-task will evaluate the models' performance on identifying *all* candidate objects referred by a given user utterance, as their canonical object IDs as defined for each scene (*e.g.* U: "*How much is that blue shirt on the hanger?*" → (object IDs of all blue shirts on hangers in the scene).

The task will provide the ground-truth bounding boxes defining each object ID to make evaluation easier. The performance will mainly be measured for its F1 score.

2.3.2 Multimodal Coreference Resolution (MM-Coref)

The goal of this task is to resolve referential mentions in user utterances to their canonical object IDs as defined for each scene. The resolving contexts can come either through (1) the dialog context (*e.g.* A: "*This shirt comes in XL and is $29.*" → U: "*Please add it to cart.*", (2) the multimodal context (*e.g.* U: "*How much is that red shirt back there?*"), or (3) both (*e.g.* U: "*How much is the one next to the one you mentioned?*").

2.3.3 Multimodal Dialog State Tracking (MM-DST)

Following the earlier challenge, the goal of the MM-DST task is to predict slots and their corresponding values grounded on the co-existing multimodal context. This requires tracking the states of multimodal objects (in addition to textual tokens) as part of dialog states. Note that this task extends the traditional notion of the unimodal dialog state tracking (DST) problem widely studied by the DSTC community.

2.3.4 Assistant Response Generation

The goal of the task is to generate Assistant responses given user utterances, ground-truth APIs and object IDs. While the assistant agent has the ground-truth meta information on each object, the referent objects need to be described *as observed and understood* by the user through the co-observed scene or the dialog context, adding an interesting challenge to the traditional response generation tasks (*e.g.* `INFORM:RECOMMEND (OBJ_ID: 3)` → A: "*I recommend the blue jacket directly behind the one I mentioned*".

In addition, with the new annotations and utterances added for SIMMC 2.1, we expect that the entries that can leverage the identified ambiguous candidate list as part of the response will achieve the best performance (*e.g.* A: "*Which of the two red jackets are you referring to - the one on the left or the one closer to you?*", as opposed to a generic response like: "*Which one are you referring to?*") – which is the main focus of this edition of the challenge.

3 Organizers

Seungwhan Moon; `shanemoon@fb.com`
Satwik Kottur; `skottur@fb.com`
Babak Damavandi; `babakd@fb.com`
Alborz Geramifard; `alborzg@fb.com`

References

[1] H. Al Amri, V. Cartillier, R. G. Lopes, A. Das, J. Wang, I. Essa, D. Batra, D. Parikh, A. Cherian, T. K. Marks, and C. Hori. Audio visual scene-aware dialog (avsd) challenge at dstc7. 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015.

[3] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[4] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017.

[5] Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*, 2018.

[6] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017.

[7] Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*, 2019.

[8] Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486*, 2020.

[9] Matthew Henderson, Blaise Thomson, and Jason D Williams. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272, 2014.

[10] Chiori Hori, Anoop Cherian, Tim K. Marks, and Florian Metze. Audio visual scene-aware dialog track in dstc8. *DSTC Track Proposal*, 2018.

[11] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[12] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*, 2019.

[13] Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. Situated and interactive multimodal conversations. *arXiv preprint arXiv:2006.01460*, 2020.

[14] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[15] Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*, 2018.